The Journal of Open Source Software

{fairmetrics}: An R package for fairness evaluation metrics with confidence intervals

Benjamin Smith[1], **Jianhui Gao**[1], **Benson Chou**[1], **and Jessica Gronsbell**[1]

1 University of Toronto

DOI:

Software

- Review
- Repository
- Archive

Submitted:
Published:

Summary

{fairmetrics} is an R package designed to evaluate the fairness of machine learning models through a range of specialized metrics for which a model can be classified as "fair". It supports fairness assessments of popular group-based criteron, such as independence, separation, sufficency and others. The package enables statistical inference on fairness metrics through calculation of bootstrap confidence intervals (CIs). In addition, {fairmetrics} offers convenient wrapper functions to compute multiple metrics simultaneously and includes datasets derived from the MIMIC-II clinical database (Goldberger et al. 2000; Raffa 2016) for illustrating its use.

Statement of Need

Machine learning (ML) offers significant potential for predictive modelling in biomedical research (Rajpurkar et al. 2022). Despite its promise, there is substantial evidence that, without appropriate forethought and planning, ML models can introduce or exacerbate health inequities by making less accurate decisions for certain groups or individuals (Grote and Keeling 2022). While existing software can compute fairness metrics, none provide out-of-the-box statistical inference, leaving practitioners without guidance on the uncertainty around those metrics. As ML becomes increasingly embedded in healthcare systems, ensuring equitable model performance across diverse populations is essential(Gao et al. 2024). The {fairmetrics} R package fills this gap by offering a suite of popular group-fairness metrics along with bootstrap-based confidence intervals, enabling more rigorous and interpretable assessments of fairness in biomedical ML.

Fairness Criteria

An ML model can be evaluated as being "fair" through three major criteron: group fairness, individual fairness and causal fairness. Group fairness deems a model fair if its predictions are similarly accurate or callibrated across a predefined set of groups, individual fairness insists that similar individuals should receive similar outcomes, and causal fairness leverages causal models that groups do not have an unjust influence on model predictions (Gao et al. 2024). {fairmetrics} focuses on calculating group fairness metrics as they are commonly used in biomedical settings. The groups in question are most often defined by protected attributes, such as age or race (Mehrabi et al. 2021).

Group fairness criteria are commonly categorized into three main types: independence, separation, and sufficiency (Barocas, Hardt, and Narayanan 2023; Berk et al. 2018;

Castelnovo et al. 2022). Independence requires that an ML model's predictions be statistically independent of the protected attribute. Separation demands that the model's predictions be independent of the protected attribute conditional on the true outcome class (i.e., within the positive and negative classes). Sufficiency requires that, given a model's prediction, the likelihood of the true outcome is independent of the protected attribute—aiming to equalize error rates across groups for similar prediction score. The {fairmetrics} package computes a range of group fairness metrics along with bootstrap-based confidence intervals. These metrics are grouped below according to the three core fairness frameworks described above.

Independence

- **Statistical Parity:** Compares the overall rate of positive predictions between groups, irrespective of the true outcome.
- **Conditional Statistical Parity:** Restricts the comparison of positive prediction rates to a specific subgroup (e.g., within a hospital unit or age bracket), offering a more context-specific fairness assessment.

Separation

- **Equal Opportunity:** Focuses on disparities in false negative rates (FNR) between two groups, quantifying any difference in missed positive cases.
- **Predictive Equality:** Compares false positive rates (FPR) between groups, ensuring that no group is disproportionately flagged as positive when the true outcome is negative.
- **Positive Class Balance:** Checks whether, among individuals whose true outcome is positive, the distribution of predicted probabilities is comparable across groups.
- **Negative Class Balance:** Checks whether, among individuals whose true outcome is negative, the distribution of predicted probabilities is comparable across groups.

Sufficiency

- **Predictive Parity:** Compares positive predictive values (PPV) across groups, assessing whether the precision of positive predictions is equivalent.

Other Criteria

- **Brier Score Parity:** Assesses whether the Brier score—the mean squared error of probabilistic predictions—is similar across groups, indicating comparable calibration.
- **Accuracy Parity:** Measures whether the overall accuracy of a predictive model is equivalent across different groups.
- **Treatment Equality:** Compares the ratio of false negatives to false positives across groups, ensuring the balance of missed detections versus false alarms is consistent.

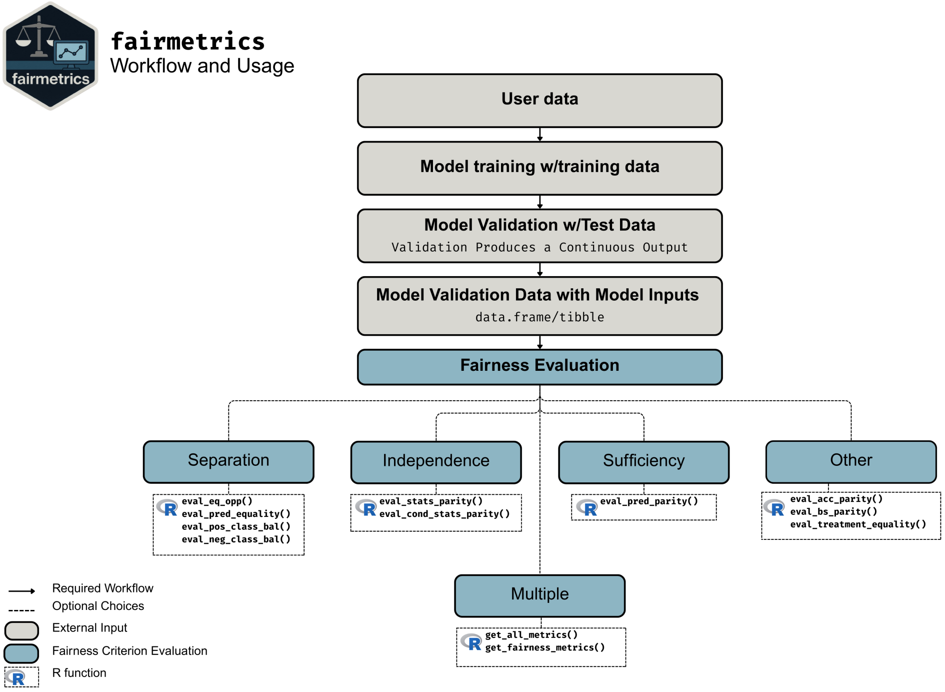

Figure 1: Workflow for using {fairmetrics} to evaluate model fairness across multiple criteria.

Evaluating Fairness Criteria

The {fairmetrics} package requires that a model has been trained and validated with data that has been split appropriately. The primary required input is a data frame or tibble which containing the model predictions, true outcomes, and protected attributes. Figure 1 shows the workflow for using {fairmetrics}. It is possible to evaluate a model for a specific or multiple group fairness metrics.

More concretely, an example of how to use the {fairmetrics} package is shown below. The example uses the `mimic_preprocessed` dataset, which is a preprocessed version of the MIMIC-II database (Goldberger et al. 2000; Raffa 2016).

Consider a random forest model trained to predict the risk of heart failure in patients. Suppose we are interested in evaluating the model's fairness with respect to patient gender with predictive equality.

```
library(fairmetrics)
library(dplyr)
library(magrittr)
library(randomForest)

# Load the example dataset
data("mimic_preprocessed")

# Split the data into training and test sets
train_data <- mimic_preprocessed %>%
  dplyr::filter(dplyr::row_number() <= 700)

test_data <- mimic_preprocessed %>%
```

```r
  dplyr::mutate(gender = ifelse(gender_num == 1, "Male", "Female")) %>%
  dplyr::filter(dplyr::row_number() > 700)

# Train a random forest model
rf_model <- randomForest::randomForest(
  factor(day_28_flg) ~ .,
  data = train_data,
  ntree = 1000
  )

# Make predictions on the test set
test_data$pred <- predict(rf_model, newdata = test_data, type = "prob")

# Evaluate predictive equality
# (Setting message=FALSE to avoid cluttering the output)

eval_pred_equality(
  data = test_data,
  outcome = "day_28_flg",
  group = "gender",
  probs = "pred",
  cutoff = 0.41,
  message = FALSE
)

#>   Metric GroupFemale GroupMale Difference  95% Diff CI Ratio 95% Ratio CI
#> 1    FPR        0.08      0.03       0.05 [0.02, 0.08]  2.67 [1.38, 5.15]
```

This case it can be seen from both the difference and ratio of the false positive rates (FPR) that the model is biased, with a 2.67 times higher chance of a false positive prediction for heart failure for females compared to males. The 95% bootstrap confidence intervals for both the difference and the ratio of the FPR are also shown, which can be used to further assess the statistical significance of the difference, confirming the initial evaluation of the FPR for female and male patients.

Related Work

Other R packages similar to {fairmetrics} include {fairness}(Kozodoi and V. Varga 2021) and {fairmodels}(Wiśniewski and Biecek 2022). The differences between {fairmetrics} and these other packages is twofold. The primary difference between is that {fairmetrics} allows for the calculation of estimated confidence intervals of fairness metrics via bootstrap, which allows for more meaningful inferences about the fairness metrics calculated. Additionally, the {fairness} package has fewer dependencies and a lower memory footprint, making the for a more environment agnostic tool that can be used with modest hardware. Table 1 shows the memory comparison when loading each library.

For python users, the {fairlearn} library (Weerts et al. 2023) provides a broader set of fairness metrics and algorithms. The {fairmetrics} package is designed for seemless integration with R workflows, making it a more convenient choice for R-based ML applications.

Package	Memory (MB)
fairmodels	17.02
fairness	117.61
fairmetrics	0.05

Table 1: Memory usage of fairmetrics vs similar packages (MB)

Licensing and Availability

The {fairmetrics} package is under the MIT license. It is available on CRAN and can be installed by using `install.packages("fairmetrics")`. A more in-depth tutorial can be accessed at: https://jianhuig.github.io/fairmetrics/articles/fairmetrics.html. All code is open-source and hosted on GitHub. All bugs and inquiries can be reported at https://github.com/jianhuig/fairmetrics/issues/.

References

Barocas, Solon, Moritz Hardt, and Arvind Narayanan. 2023. *Fairness and Machine Learning: Limitations and Opportunities*. Cambridge, Massachusetts: The MIT Press.

Berk, Richard, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2018. "Fairness in Criminal Justice Risk Assessments: The State of the Art." *Sociological Methods & Research* 50 (1): 3–44. https://doi.org/10.1177/0049124118782533.

Castelnovo, Alessandro, Riccardo Crupi, Greta Greco, Daniele Regoli, Ilaria Giuseppina Penco, and Andrea Claudio Cosentini. 2022. "A Clarification of the Nuances in the Fairness Metrics Landscape." *Scientific Reports* 12 (1). https://doi.org/10.1038/s41598-022-07939-1.

Gao, Jianhui, Benson Chou, Zachary R. McCaw, Hilary Thurston, Paul Varghese, Chuan Hong, and Jessica Gronsbell. 2024. "What Is Fair? Defining Fairness in Machine Learning for Health." *arXiv.org*. https://arxiv.org/abs/2406.09307.

Goldberger, Ary L., Luis A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley. 2000. "PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals." *Circulation [Online]* 101 (23): e215–20. https://doi.org/10.1161/01.CIR.101.23.e215.

Grote, Thomas, and Geoff Keeling. 2022. "Enabling Fairness in Healthcare Through Machine Learning." *Ethics and Information Technology* 24 (3): 39. https://doi.org/10.1007/s10676-022-09658-7.

Kozodoi, Nikita, and Tibor V. Varga. 2021. *Fairness: Algorithmic Fairness Metrics*. https://CRAN.R-project.org/package=fairness.

Mehrabi, Ninareh, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. "A Survey on Bias and Fairness in Machine Learning." *ACM Computing Surveys* 54 (6): 115:1–35. https://doi.org/10.1145/3457607.

Raffa, Jesse. 2016. "Clinical Data from the MIMIC-II Database for a Case Study on Indwelling Arterial Catheters (Version 1.0)." https://doi.org/10.13026/C2NC7F. https://doi.org/10.13026/C2NC7F.

Rajpurkar, Pranav, Emma Chen, Oishi Banerjee, and Eric J. Topol. 2022. "AI in Health and Medicine." *Nature Medicine* 28 (1): 31–38. https://doi.org/10.1038/s41591-021-01614-0.

Weerts, Hilde, Miroslav Dudík, Richard Edgar, Adrin Jalali, Roman Lutz, and Michael Madaio. 2023. "FairLearn: Assessing and Improving Fairness of AI Systems." *arXiv.org*. https://arxiv.org/abs/2303.16626.

Wiśniewski, Jakub, and Przemysław Biecek. 2022. “Fairmodels: A Flexible Tool for Bias Detection, Visualization, and Mitigation in Binary Classification Models.” *The R Journal* 14 (1): 227–43. https://doi.org/10.32614/RJ-2022-019.